Washington D.C.
October 25, 2005

Semi-Automatic Indexing of Full Text Biomedical Articles

Clifford W. Gay

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Acknowledgments

Alan R. Aronson, PhD.

Mehmet Kayaalp, M.D., PhD.

Outline

Introduction

- The System: Medical Text Indexer (MTI)
- The Data: Online biomedical journals
- The Task: Emulate Medline indexing using full text

Results

- Observations on PubMed Central articles
- Model selection results
- Recent work

Introduction

The System: Medical Text Indexer (MTI)

The Data: Online medical journals

The Task: Emulate Medline indexing using full text

Results

Observations on PubMed Central articles

Model selection results

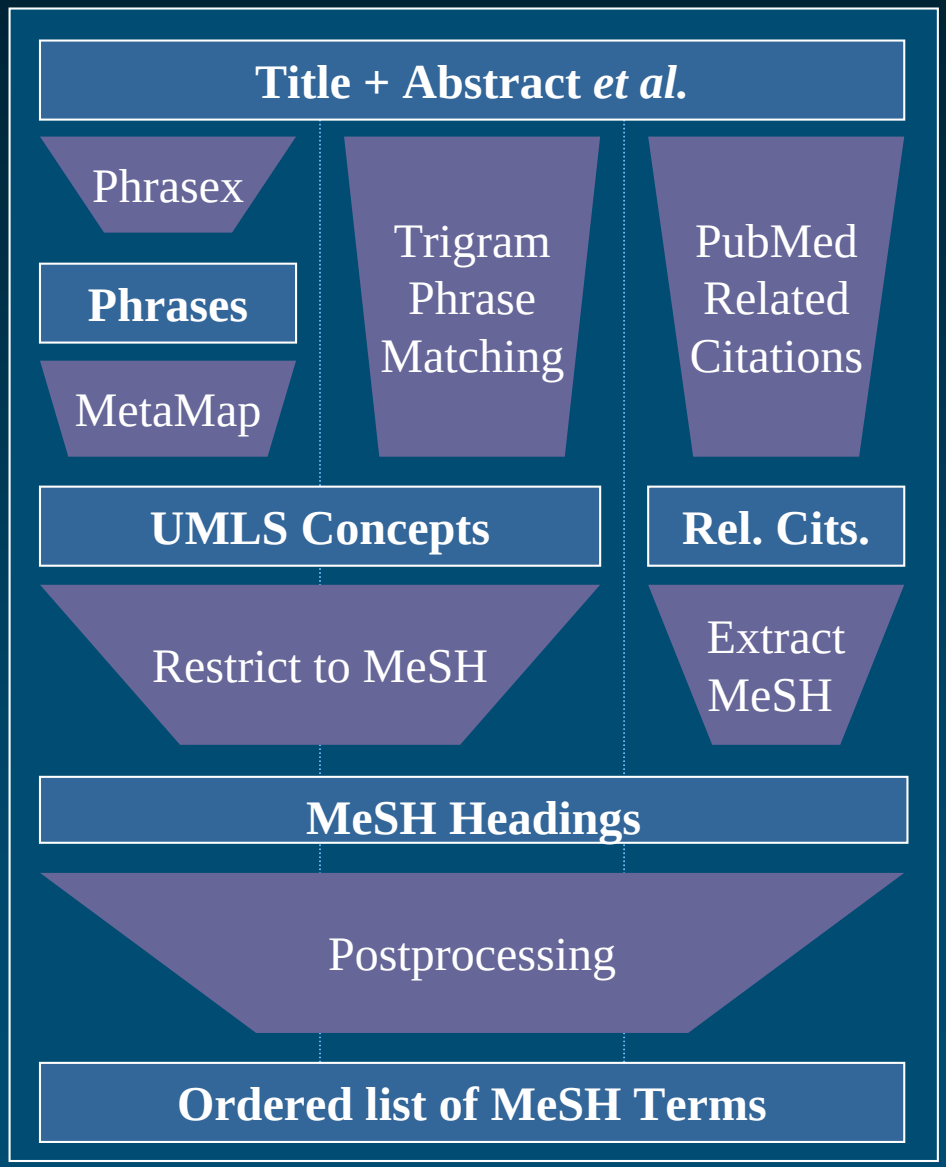
Recent work

Why Semi-Automatic Indexing?

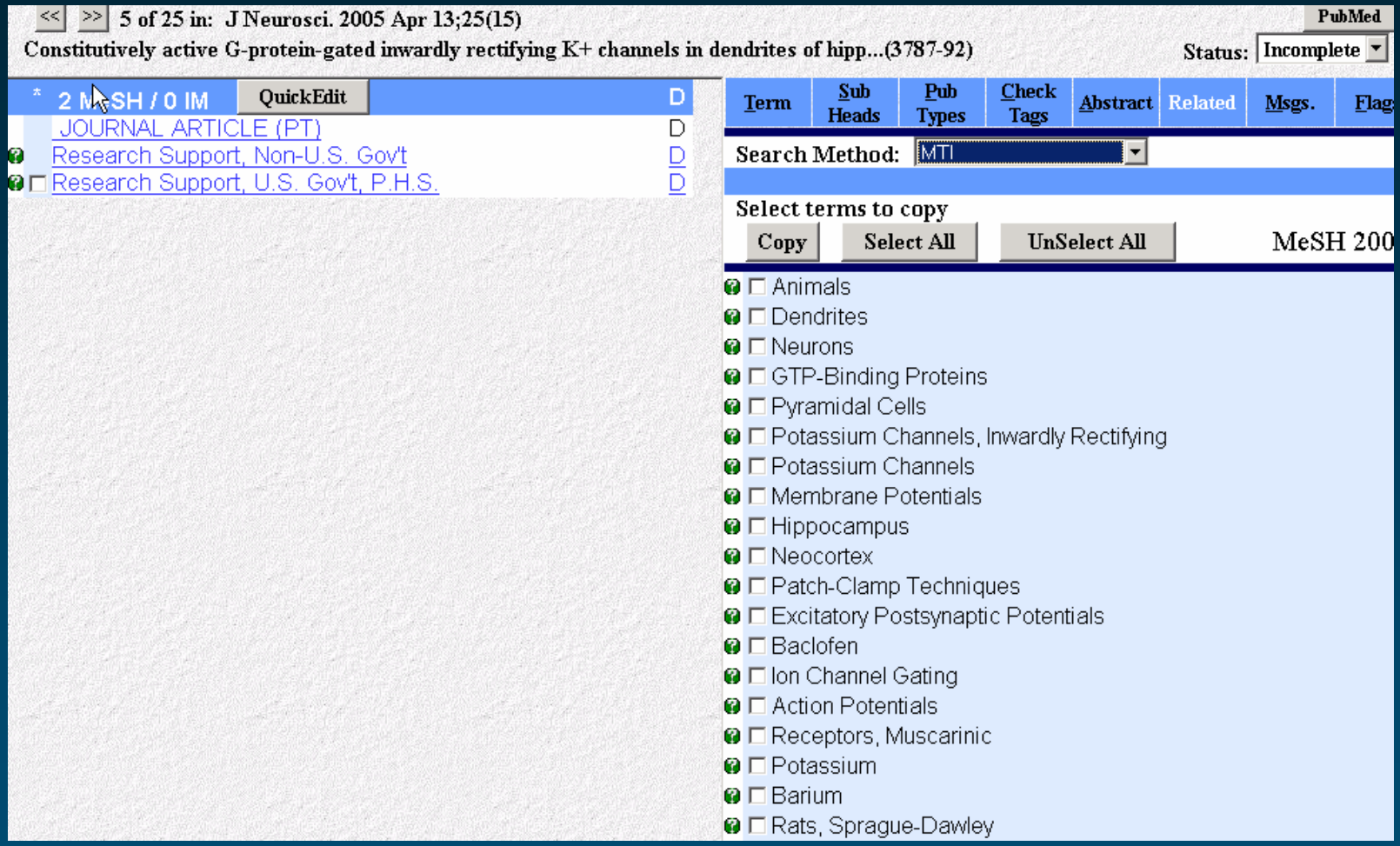
- U.S. National Library of Medicine indexes 5000 journal titles
  - Supports over 60 million PubMed searches each month
  - Has 130 indexers
  - Indexed 570,000 articles in 2004
    - Will need to index 1,000,000 very soon
  - Automated support is helping to meet this demand
    - MTI was used on 26% of articles in 2004
- More about MTI

  Aronson AR, **Mork JG**, **Gay CW**, **Humphrey SM**, **Rogers WJ**.
  The NLM Indexing Initiative's Medical Text Indexer.
  Medinfo. 2004; 11(Pt 1): 268-72. PMID: 15360816

Medical Text Indexer (MTI)



DCMS with MTI Suggestions



Introduction

The System: Medical Text Indexer (MTI)

The Data: Online biomedical journals

The Task: Emulate Medline indexing using full text

Results

Observations on PubMed Central articles

Model selection results

Recent work

Why Full Text?

- Medical Text Indexer uses article title and abstract
- However
  - Human indexers taught not to use abstract
  - Author's complete intent may not be in abstract
  - Check tags may only appear in a table or methods section.
- If MTI indexes from full text articles it may
  - Find central concepts missing from abstract
  - Identify terms when article has no abstract
  - More accurately select check tags
  - Be in better compliance with indexing policy

Test Collection Selection

- Available online from PubMed Central
- Consistent XML format
  - Identifies title, abstract, sections, tables, figures, references, etc.
- 500 articles from 17 diverse biomedical journals
- Did not use:
  - References
  - Graphics
  - Math

Test Collection

5 Clinical journals (165):

- Breast Cancer Research (11)
- Journal of Clinical Microbiology (80)

3 Organization based journals (28):

- Journal of American Medical Informatics Assoc. (10)
- Proceeding of the National Academy of Sciences (11)

9 Journals in other categories:

- Pharmacology (65); Biochemistry (65); Plants (46); Molecular Biology (45); Learning (30); Hospitals (22)

Introduction

The System: Medical Text Indexer (MTI)

The Data: Online medical journals

The Task: Emulate Medline indexing using full text

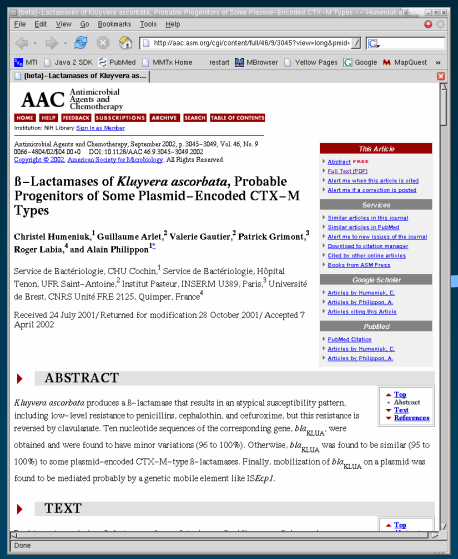
Results

Observations on PubMed Central articles

Model selection results

Recent work

Indexing Task

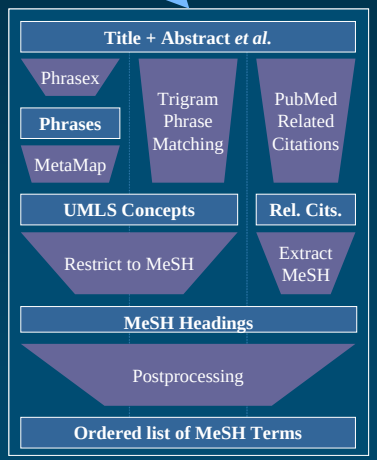


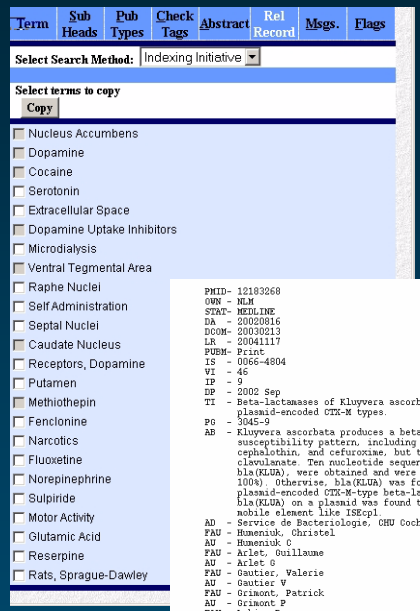
ß-Lactamases of *Kluyvera ascorbata*, Probable Progenitors of Some Plasmid-Encoded CTX-M Types

Christel Humeniuk,[1] Guillaume Arlet,[2] Valerie Gautier,[2] Patrick Grimont,[3] Roger Labia,[4] and Alain Philippon[1]*

ABSTRACT

TEXT






```
PMID- 12183268
OWN - NLM
STAT- MEDLINE
DA  - 20020816
DCOM- 20030213
LR  - 20041117
PUBM- Print
IS  - 0066-4804
VI  - 46
IP  - 9
DP  - 2002 Sep
TI  - Beta-lactamases of Kluyvera ascorbata, probable progenitors of some
      plasmid-encoded CTX-M types.
PG  - 3045-9
AB  - Kluyvera ascorbata produces a beta-lactamase that results in an atypical
      susceptibility pattern, including low-level resistance to penicillins,
      cephalothin, and cefuroxime, but this resistance is reversed by
      clavulanate. Ten nucleotide sequences of the corresponding gene,
      bla(KLUA), were obtained and were found to have minor variations (96 to
      100%). Otherwise, bla(KLUA) was found to be similar (95 to 100%) to some
      plasmid-encoded CTX-M-type beta-lactamases. Finally, mobilization of
      bla(KLUA) on a plasmid was found to be mediated probably by a genetic
      mobile element like ISEcp1.
AD  - Service de Bacteriologie, CHU Cochin, Paris, France.
FAU - Humeniuk, Christel
AU  - Humeniuk C
FAU - Arlet, Guillaume
AU  - Arlet G
FAU - Gautier, Valerie
AU  - Gautier V
FAU - Grimont, Patrick
AU  - Grimont P
FAU - Labia, Roger
AU  - Labia R
FAU - Philippon, Alain
AU  - Philippon A
LA  - eng
SI  - GENBANK/AF252622
SI  - GENBANK/AF252623
SI  - GENBANK/AF311345
SI  - GENBANK/AJ272538
PT  - Journal Article
PL  - United States
TA  - Antimicrob Agents Chemother
JID - 0315061
RN  - 0 (Plasmids)
RN  - EC 3.5.2.6 (beta-Lactamases)
RN  - EC 3.5.2.6 (beta-lactamase bla(KLUA), Klyvera ascorbata)
SB  - IM
MH  - Enterobacteriaceae/drug effects/*enzymology/genetics
MH  - Genes, Bacterial/genetics
MH  - Genotype
MH  - Kinetics
MH  - Microbial Sensitivity Tests
MH  - Molecular Sequence Data
MH  - Plasmids/*genetics
MH  - Research Support, Non-U.S. Gov't
MH  - beta-Lactamases/*genetics/*metabolism
EDAT- 2002/08/17 10:00
MHDA- 2003/02/14 04:00
PST - ppublish
SO  - Antimicrob Agents Chemother 2002 Sep;46(9):3045-9.
```

Example Article

Medline Indexing

beta-Lactamases
/*genetics /*metabolism
Enterobacteriaceae/drug effects
/*enzymology/genetics
Plasmids/*genetics
Genes, Bacterial/genetics
Genotype
Kinetics
Microbial Sensitivity Tests
Molecular Sequence Data
Research Support, Non-U.S. Gov't

MTI Indexing

- beta-Lactamases
- Plasmids
- Enterobacteriaceae
- beta-Lactam Resistance
- Conjugation, Genetic
- Cephalosporin Resistance
- Cefotaxime
- Nucleotide Sequences
- Molecular Sequence Data
- Cephalosporins
- Chromosomes, Bacterial
- DNA, Bacterial
- DNA Transposable Elements
- Escherichia coli
- Genes, Bacterial
- Cloning, Molecular
- Klebsiella pneumoniae
- Amino Acid Sequence
- Microbial Sensitivity Tests
- Cephalothin
- Proteus mirabilis
- Erwinia
- Salmonella typhimurium
- Enterobacteriaceae Infections
- Lactams

MMI REL MMI & REL

Recall = 0.67 Precison = 0.24 F_2 measure = 0.492

Evaluation

F_2 Measure

- Weighted harmonic mean of Recall and Precision
- Weights Recall twice as important as Precision
- Values: 0.0 to 1.0

Computed for each article and averaged

Introduction

The System: Medical Text Indexer (MTI)

The Data: Online medical journals

The Task: Emulate Medline indexing using full text

Results

Observations on PubMed Central articles

Model selection results

Recent work

Section Header Classes

Semantically equivalent section headers

MATERIALS AND METHODS class:

- Materials and Method(s)
- Method(s)
- Scoring Methods
- Experimental Procedures
- Other Methods Tested

CAPTIONS class:

- the titles and captions from tables and figures

Section Class Performance

Section Class	Average F_2
CAPTIONS	0.3175
ABSTRACT	0.2960
INTRODUCTION	0.2869
RESULTS	0.2790
DISCUSSION	0.2734
NO HEADER	0.2574
…	…
CONCLUSIONS	0.1961
ABBREVIATIONS	0.1304

Introduction

The System: Medical Text Indexer (MTI)

The Data: Online medical journals

The Task: Emulate Medline indexing using full text

Results

Observations on PubMed Central articles

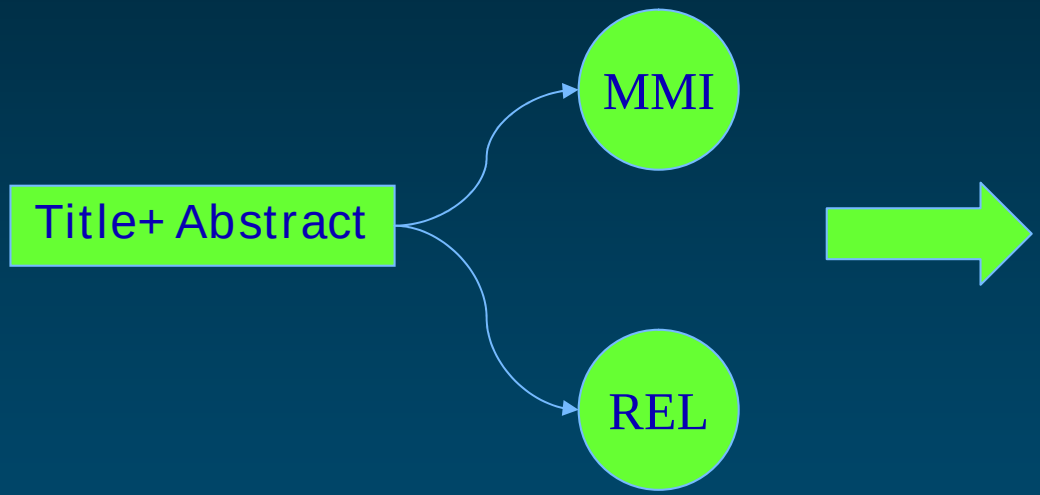
Model selection results

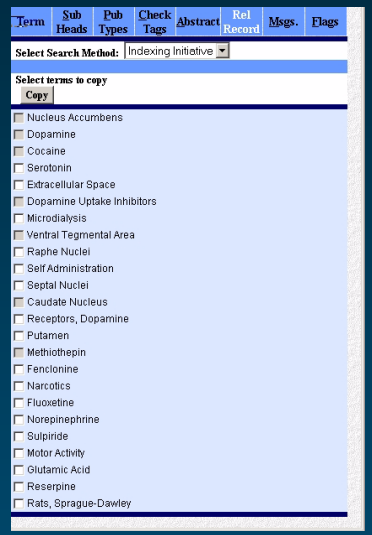
Recent work

Experiments

- Varied MTI components used
  - MetaMap Indexing (MMI)
  - Related Citations (REL)
- Varied section classes processed
  - Used model selection
  - Used binary weighting for sections
- A model is
  - A selection of section classes and
  - The text in those sections
  - That represents the article

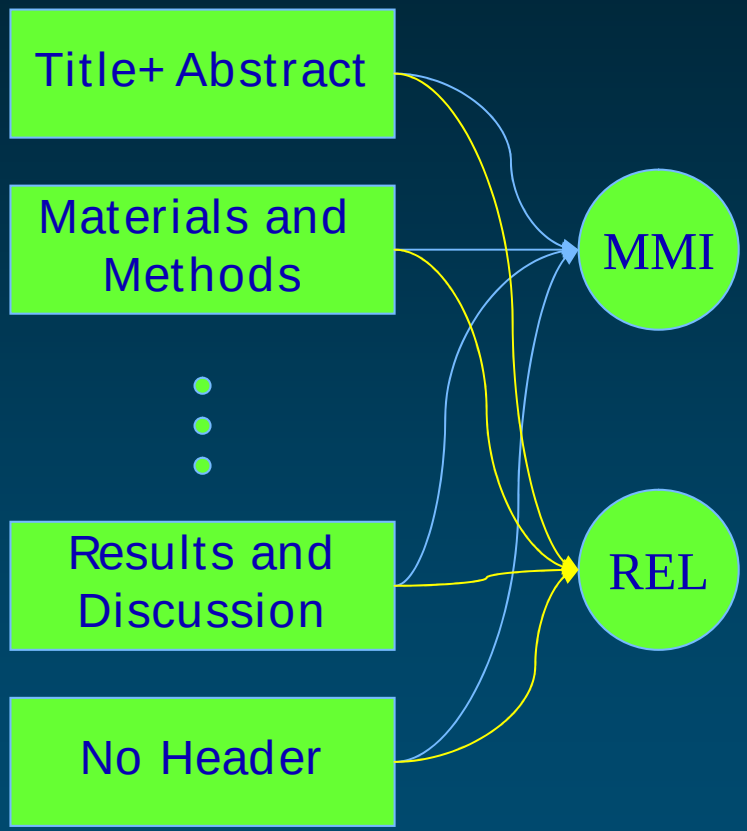
Production Baseline



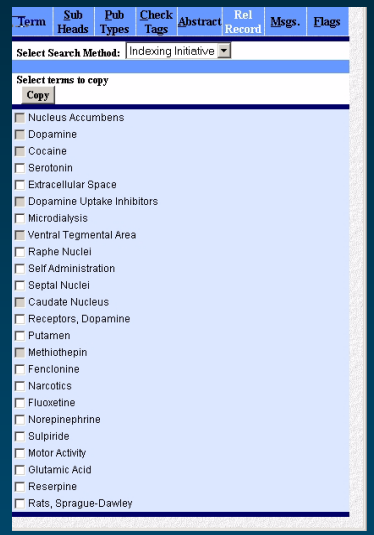


$F_2 = 0.457$

Naive Mode

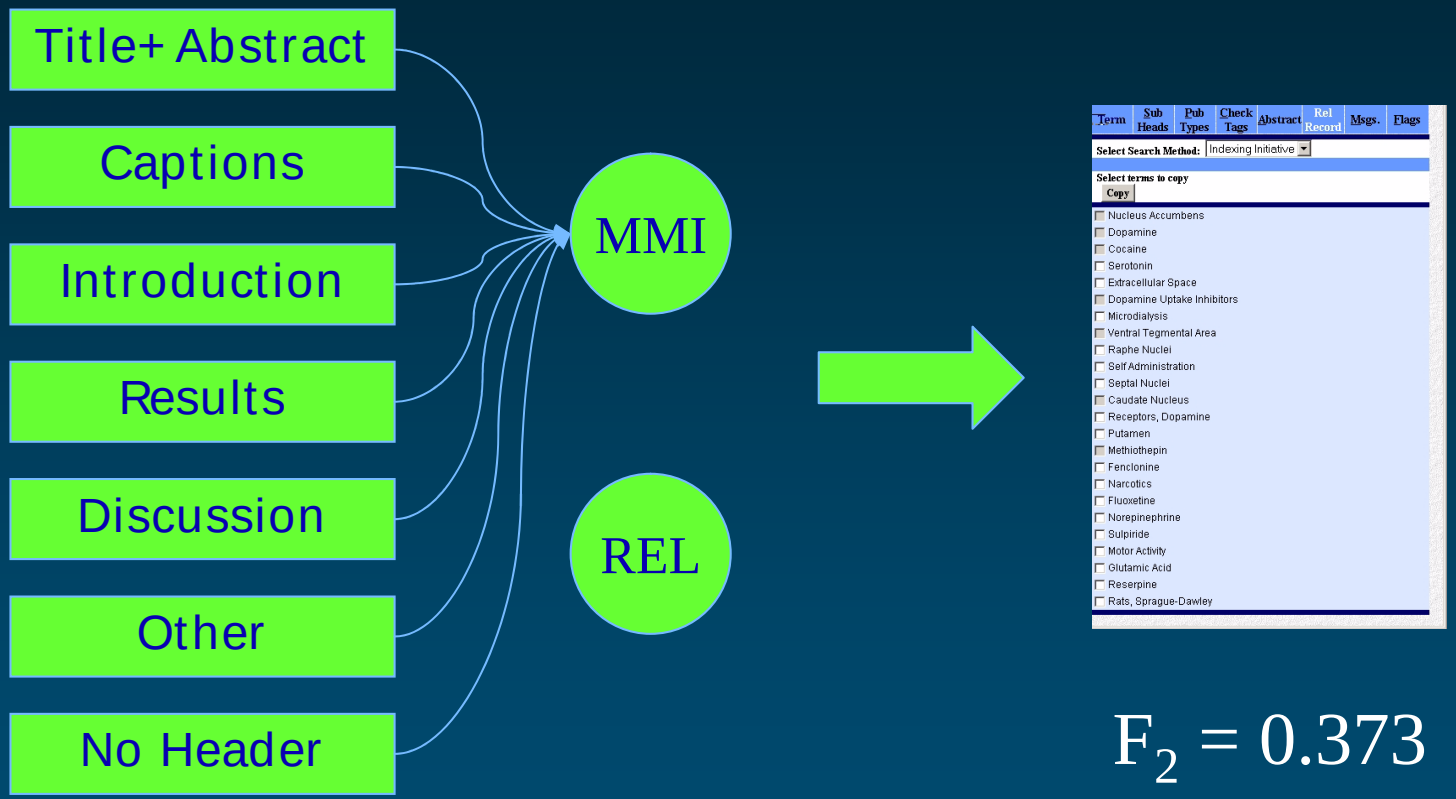


All Section Classes



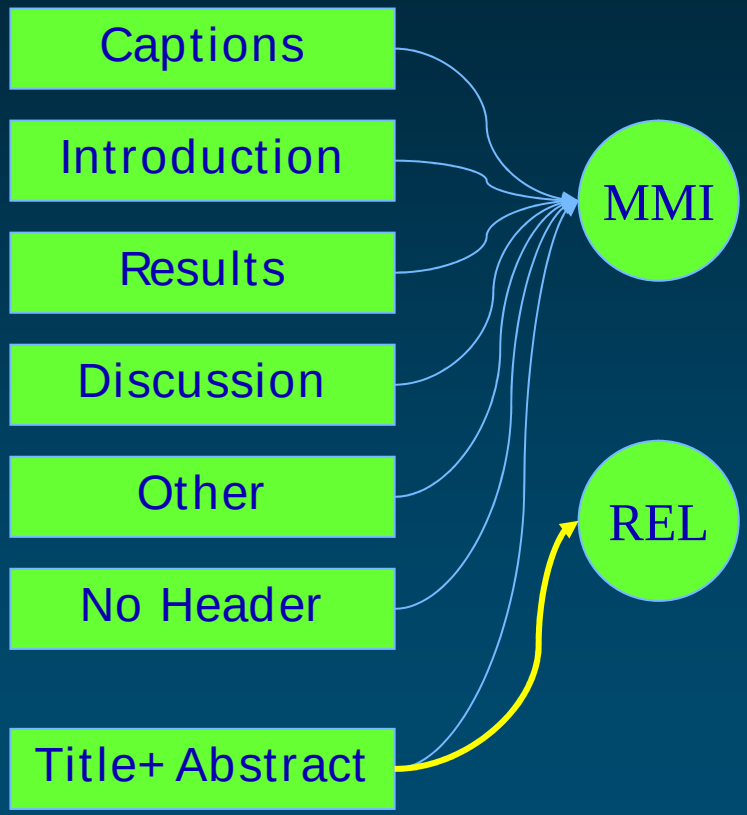
$F_2 = 0.453$
(- 0.9%)

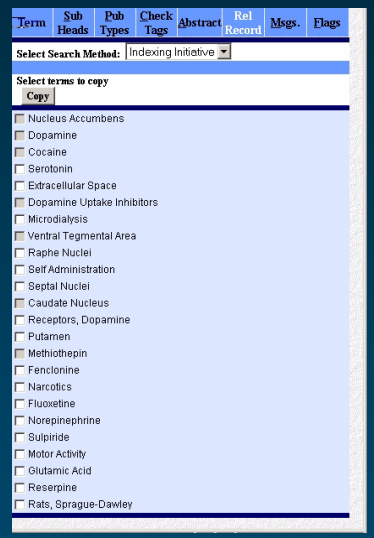
MetaMap Indexing Mode



$F_2 = 0.373$
(-18.4%)

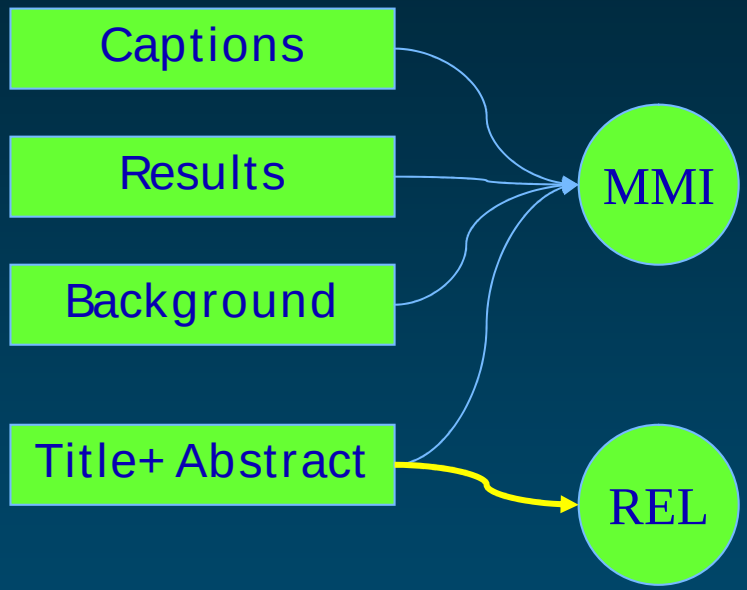
Augmented Mode

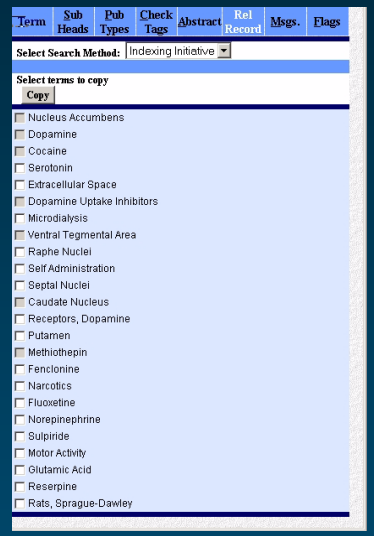




$F_2 = 0.475$
(+3.9%)

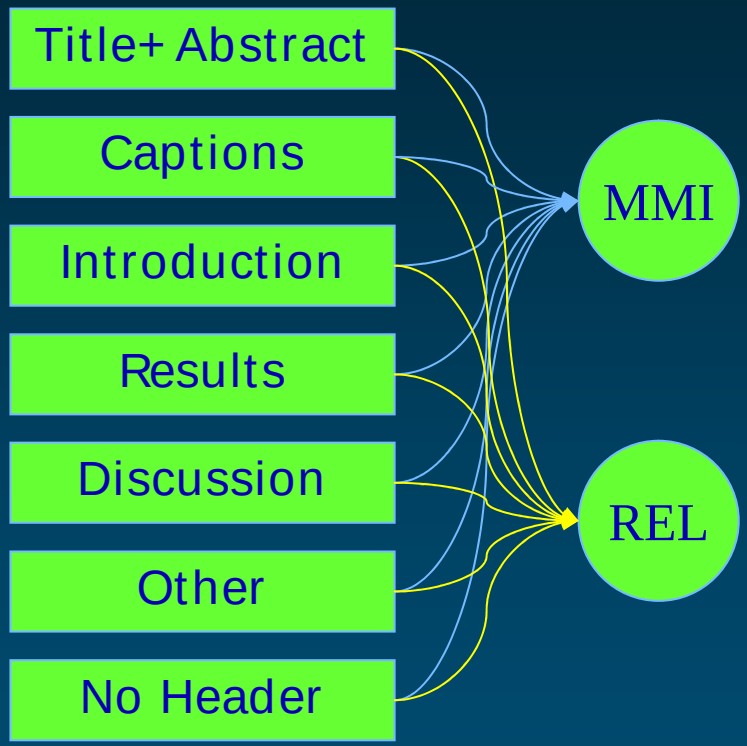
Refined Augmented Mode



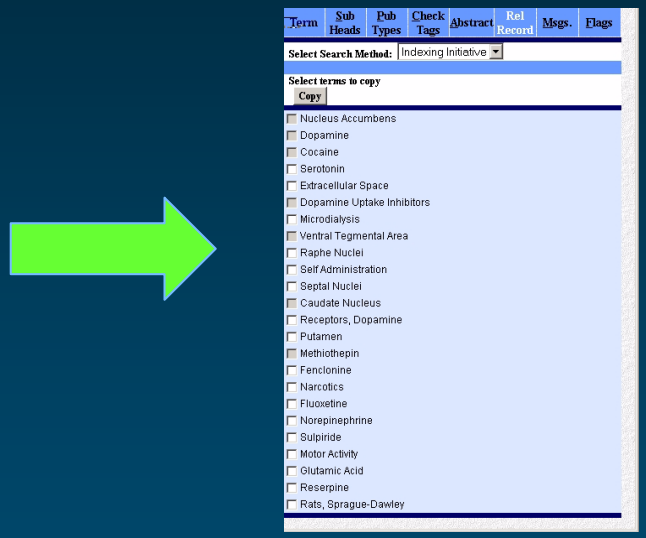


$F_2 = 0.485$
(+ 6.1%)

Full MTI Mode

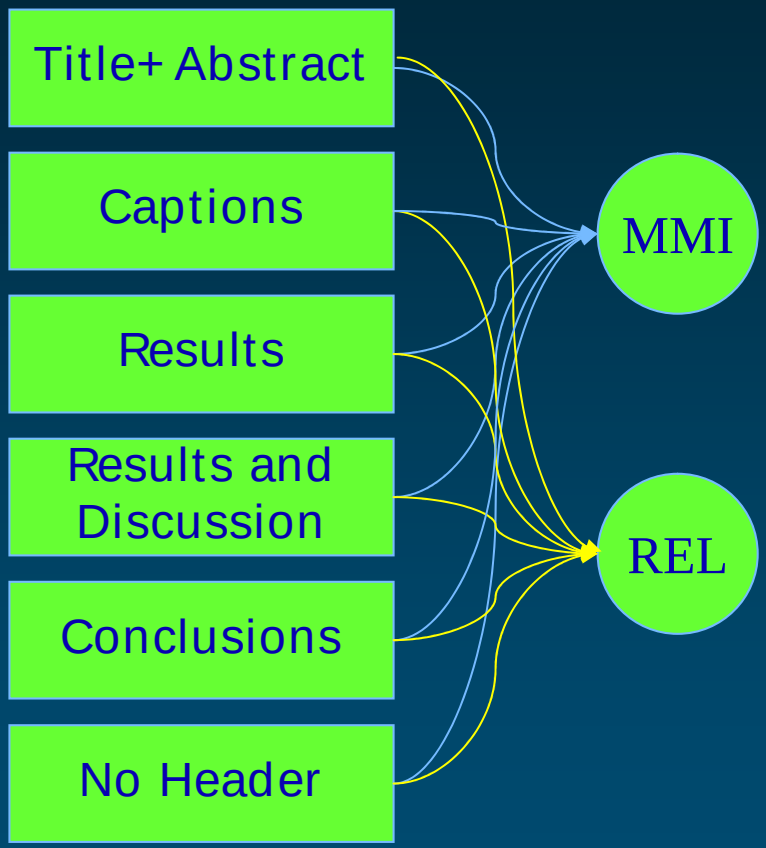


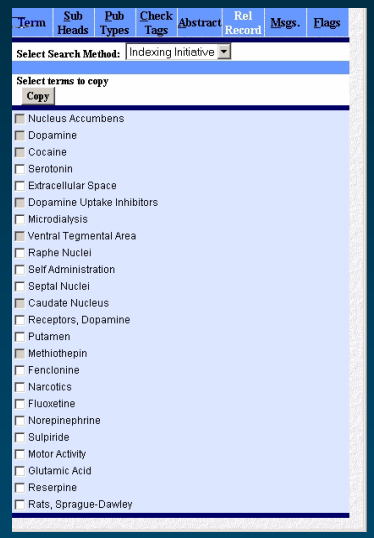
MMI model



$F_2 = 0.488$
(+ 6.8%)

Refined Full MTI





$F_2 = 0.491$
(+ 7.4%)

MTI Performance Summary

Indexing Model	Recall	Precision	Avg. F_2
Production Baseline (Ti, Ab)	**0.53**	**0.32**	**0.457**
Naive Mode (full text)	0.57	0.27	0.453
Augmented Mode (MMI + REL (Ti, Ab))	0.59	0.29	0.475
Augmented Mode (refined)	0.60	0.30	0.485
Full MTI (MMI + REL common sections)	0.60	0.30	0.488
Full MTI (refined)	**0.60**	**0.31**	**0.491**

Introduction

The System: Medical Text Indexer (MTI)

The Data: Online medical journals

The Task: Emulate Medline indexing using full text

Results

Observations on PubMed Central articles

Model selection results

Recent work

Improvement Potential

With current model

- No cut off at 25 terms yields maximum recall of 0.79

If all good terms prioritized correctly

- $F_2 = 0.64$
- Improvement over baseline
  - 7% → 40%

Increase REL Citations

MTI currently uses 10 Related Citations

Optimal number for full text articles is 15

Best model confirmed for this setting

Additional Improvement in F_2 = 0.01

Summarization

- Selecting important text before MTI processing
- Using Yeh, Ke, Yang, Meng approach
- Combines
  - Latent Semantic Analysis and
  - Salton's Text Relationship Map
- Start with current model
- Document representation includes
  - Bag of words
  - MetaMap identified concepts

NLM Indexing Initiative

Contact: cliff@nlm.nih.gov
Web: ii.nlm.nih.gov/fulltext.shtml

Clifford W. Gay

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

NONE Sections

- Most appear in articles that have no abstract
  - 20/23
- Some are errors
  - 4 have “Introduction” header in publisher version
  - 2 appear within other sections with headers.
- Many contain the primary text of the article
  - Comments, Editorials, Letters (11/23)

Other Sections

Other section class has 525 sections (16%)

Non-standard article organization

- Common in Review articles

Example

- **ß-Lactamases of *Kluyvera ascorbata*, Probable Progenitors of Some Plasmid-Encoded CTX-M Types**
  - **Bacterial strains.**
  - **Antimicrobial agents and susceptibility testing.**
  - **Kinetic and IEF analyses.**
  - **Genetic characterization of *bla*KLUA.**
  - **Genetic environment of *bla*KLUA-1.**
  - **Arguments for mobilization of chromosomal *bla*KLUA gene.**

Ranking Function

Made ranking function for Related Citations more like MetaMap Indexing.

Resulted in a more inclusive model

- Materials and Methods
- Introduction

F2 measure = 0.4865

Tuning Path Weight

Ratio of weights between the two indexing paths

- MetaMap Indexing – 7
- Related Citations – 2

No improvement possible

Partial Weight for Singleton Headers

- OTHER section class
  - Header is unique
  - Contain content terms
- Gave section class weight between 0 and 1
  - Some recall improvement
  - No collection wide improvement in F_2